DOCUMENT RESUME

ED 189 130 TM 800 296

AUTHOR Mills, Craig N.; Hambleton, Ronald K.
TITLE Guidelines for Reporting Criterion-Referenced Test Score Information.
PUB DATE Apr 80
NOTE 47p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Educational Objectives; Elementary Secondary Education; *Guidelines; *Information Needs; Information Utilization; *Scores; *Testing Problems; *Test Interpretation; Test Items
IDENTIFIERS *Test Reporting

ABSTRACT

General guidelines exist for reporting and interpreting test scores, but there are short comings in the available technology, especially when applied to criterion-referenced tests. Concerns that have been expressed in the educational measurement literature address the uses of test scores, the manner of reporting scores, limited testing knowledge among users, presentation of results to parents and students, and use of computer technology to report test scores. Several activities must occur before high quality test score reports can be prepared. These activities include the specification of information needs, building a testing program consistent with needs, identification of audiences and their levels of testing knowledge, proper test selection, and proper test construction. A rating system which can be used in designing or evaluating criterion-referenced test score reporting systems is presented, based on a logical analysis of criterion-referenced tests; current uses of the tests; and information needs of parents and students, building administrators, and district administrators. This rating system is organized around seven major categories of concern: information about objectives, information at the item level, information at the objective level, information at the subtest level, subject summaries, specialized services, and general services. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. Mills
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED189130

Guidelines for Reporting Criterion-Referenced Test Score Information

Craig N. Mills and Ronald K. Hambleton
University of Massachusetts, Amherst

Although research during the last several years has led to the development of methods for the construction and validation of criterion-referenced tests, little attention has been paid to the manner in which test score information should be reported. While the likely explanation is that measurement specialists have spent their time researching topics which logically precede the reporting and interpreting of test scores (for example, preparing methods for assessing validity, assessing test score reliability, setting cut-scores and determing test lengths), it is unfortunate that reporting test score information has not received more attention. The purpose of a testing program is, after all, to provide usable information to appropriate audiences in a convenient format.

The purposes of this paper are to (1) review relevant literature in relation to reporting and utilizing criterion-referenced test and norm-referenced test results, (2) describe desirable characteristics of test score reports, and (3) provide guidelines and a system for use in the design and evaluation of criterion-referenced test score reports for school settings. The system allows practitioners to specify a list of information needs which apply in their situation. The list can then be used to design a reporting system or to evaluate the reporting systems which accompany commercially available criterion-referenced tests.

TM800296

Guidelines for Reporting Criterion-Referenced Test Score Information[1,2]

Craig N. Mills and Ronald K. Hambleton
University of Massachusetts, Amherst

Research during the past several years has led to the development of methods for the preparation of criterion-referenced tests and for the validation of test score uses (Hambleton & Eignor, 1979; Popham, 1978a). Criterion-referenced test score reporting is a topic, however, that has received only scant attention in the educational measurement literature. For example, in two recent reviews of the criterion-referenced testing field (Hambleton, Swaminathan, Algina, & Coulson, 1978; Popham, 1978a) only a few sentences in each dealt with test score reporting. A likely explanation is that measurement specialists have spent their research time on topics which logically precede test score reporting (for example, test development, assessment of test score reliability and validity, determination of test lengths, and standard-setting).

It is unfortunate that criterion-referenced test score reporting has received so little attention because the purpose of a testing program is, after all, to provide usable test score information in a convenient format. Test score information that is inappropriate, confusing, or in any other way unsuited to the needs of potential test score users will be of limited value.

---

[1]Laboratory of Psychometric and Evaluative Research Report No. 100. Amherst, MA: School of Education, University of Massachusetts, 1979.

[2]A paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

The quality and appropriateness of criterion-referenced test score reporting impacts directly on the use of the test score information. Test score results are used to make decisions that range from diagnosis of learning deficiencies and monitoring student progress in objectives-based programs to program evaluation and funding decisions. Since many decisions have long-term implications for examinees, it is imperative that the information provided to decision-makers be the appropriate type and that it be in a format which facilitates effective decision-making.

One might be tempted to suggest that reporting methods developed over the years for use with norm-referenced test scores, could, with minor revisions, provide the needed technology. However, the nature of the statements to be made about examinees based on their criterion-referenced test performance are substantially different. Norm-referenced tests (NRTs) are constructed, principally, to facilitate comparisons among individuals (or groups) in relation to the performance of a norm group. Criterion-referenced tests (CRTs), on the other hand, are constructed to facilitate the interpretation of individual (or group) test performance in relation to a set of well-defined objectives or competencies (Hambleton & Eignor, 1978). Criterion-referenced test scores are used to make inferences about examinee performance in domains of content describing objectives included in the test and to make a variety of mastery/non-mastery decisions. Therefore, the nature of NR and CR reporting systems will be different to reflect the principal types of information being reported. However, many of the characteristics of a good reporting system will be common to both types of tests. And, since NRT reporting

systems have been around for a lot longer, there are things that can be learned from these systems that will impact positively on CRT score reporting systems.

The areas of CRT score reporting and utilization require study since (1) little direct research has been carried out, (2) NRT score reporting technology is helpful but it does not address all of the problems encountered in reporting CRT scores, and (3) there is considerable interest and activity in schools around the country in using CRTs. The purposes of this study are to (1) review the relevant literature in relation to reporting and utilizing CRT and NRT results, (2) describe desirable characteristics of test score reports, and (3) provide guidelines and a system for use in the design and evaluation of test score reports for school settings.

The remainder of this paper is divided into four sections. Necessary antecedents to the preparation of high quality reports are considered in one section. The other three sections correspond to the three purposes outlined above.

Review of Literature

Since proper score reporting and interpretation is so important, test developers and practitioners need to have at their disposal, standards by which they can develop and/or evaluate test score reports. General guidelines exist for reporting and interpreting test scores and can be found in most educational measurement textbooks. But, there are shortcomings in the available technology and its application. According to Page (1977), "On one general problem of testing we can be in fair agreement: A great gap exists between the expertise of test development and the amateurish use of test scores" (p. 8-9). Fisher (1978) noted that there is no shortage of trained personnel to handle technical problems in the areas of test development, administration, and scoring. But, according to Fisher, "The problems begin when these results are communicated to various audiences, and the problems get serious when someone attempts to assign meaning to the data" (p. 35). Popham (1978b) praised many aspects of the reporting system of the California Assessment Program but observed "a certain half-heartedness in the explanatory documents that accompany CAP results" (p. 20). Hagen (1977) also expressed concern:

> Those of us who have devoted our professional lives to testing and evaluation need to pay much more attention to translating test scores into constructive actions. Many of us have been too concerned with the predictive validity of tests and have been too little concerned with what test scores mean in terms of behavior and constructive actions to be taken to facilitate the development of the individual. We need to work more closely with teachers and educators to determine what information they need in order to make education more effective and help them get this information in a form which is useful to them. (p. 167)

The sample of comments offered above does reveal, rather clearly, that the areas of test score reporting and of test score use have problems associated with them. Concerns that have been expressed in the educational measurement literature can be organized into five categories:

1. Uses of test scores,
2. Manner of reporting scores,
3. Limited testing knowledge among users,
4. Presentation of results to parents and students,
5. Use of computer technology to report test scores.

In the remainder of this review of literature each of the five areas will be briefly considered.

Uses of Test Scores

There are six major uses of test results: selection, classification, diagnosis, research, program evaluation, and description. The stated purposes of a testing program will determine to a large extent the use of the scores. However, the audience for which the scores are intended is also an important consideration. Audiences which often require information are students and their families, teachers, school administrators, and district administrators, Although some types of information may be desired by all groups, each group will have unique information needs as well.

Students and their families need test information for diagnostic uses as well as for achievement determination (Stetz, 1978). For example, test results may be used (1) to choose appropriate curricula or to make educational and vocational choices (Goslin, 1967; Kirby, Culp,

& Kirby, 1973), (2) to identify an individual's strengths and weaknesses and to develop strategies for improvement or remediation (Bradley, 1978; Goslin, 1967; Hagen, 1977), and (3) to determine achievement status during a school year (Goslin, 1967), relative achievement levels in serveral subjects areas, or performance in one subject or area over time (Gardner, 1977).

Teachers may use test results as aids in making decisions about students and in evaluating their instruction. Examples of decisions about students are decisions about grouping and placement (Wahlstrom, Danley & Raphael, 1977) and group and individual diagnosis (Rost, 1973). Evaluation of instruction includes teacher self-assessment (Wahlstrom, Regan & Jones, 1978), curricular reform (Rost, 1973; Wahlstron, Danley & Raphael, 1977) and the appropriateness of the difficulty of course objectives (Wahlstrom, Regan & Jones, 1978).

At the local level a variety of administrative uses of test results are possible. Test results can be used to compare the performance of a school or system to national norms (Gardner, 1977; Wahlstrom, Danley & Raphael, 1977). Comparisons among schools within a system may help identify patterns of achievement over time (Gardner, 1977) or to identify problem schools which may need additional resource personnel (Wahlstrom, Danley, & Raphael, 1977). Other administrative uses of test results include program evaluation and curriculum development (Goslin, 1967; Lawson & Ward, 1976; Stetz, 1978; Wahlstrom, Danley & Raphael, 1977) and the evaluation of teacher effectiveness (Goslin, 1967).

Manner of Reporting Scores

The types of statements made from test results is dependent upon the strategy utilized for measuring achievement. Ahmann (1978) and Millman (1978) list three strategies: item-centered, objective-centered, and subtest-centered. All three strategies are used with varying degrees of success with NRTs and CRTs.

In an item-centered approach, information about examinee performance on each item in a test is presented. Such a strategy can be employed to provide information about group performance on specific skills (reflected by single test items). It is not usually advisable, however, to make statements about individual examinee performance on an objective on the basis of performance on a single test item due to the unreliability of information provided by a single test item.

Objective-centered reporting involves providing an individual's or group's performance on an objective on the basis of several items which measure that objective. A test may include many objectives and have many items per objective. Multiple test items/objective increases the reliability of reported examinee objective scores, but decreases the breadth of coverage of a test, unless testing time is increased.

Subtest-centered reporting usually involves a small number of skills but many items measuring each skill. The extent to which these skills are defined will vary from one test to another. In theory, the skills should be more thoroughly defined with CRTs because of the intended uses of the test scores. The scores of a typical norm-referenced test (mathematics concepts, problem-solving, and computations; vocabulary, reading comprehension) are examples of subtest-centered reports.

A fourth strategy also exists. Millman (1970) suggests that two lists of objectives may be useful. One list would be for teachers. This list would include specific objectives to be measured. The second list, which would be for parents, would include broader objectives. Millman suggests that teachers receive information about performance on each objective whereas parents could receive information about clusters of related objectives. For example, Millman suggests that teachers may need information on objectives such as "identification of coins and converting coins to equivalent amounts of other coin values." Parents, on the other hand, might be confused by data on a large number of small objectives. It would be better to provide them with information such as "understands the dollar value of money" (p. 227).

Limited Testing Knowledge Among Users

One reason for teachers misusing or misinterpreting test scores is because they are unfamiliar with the field of tests and measurements. Goslin (1967) found that less than 40 percent of all teachers have had more than one course in educational testing. Many teachers have had no exposure to test and measurement techniques in either formal classes or in-service training. It is not surprising therefore to find that many teachers do not understand or properly use test results. The call for in-service and pre-service training to upgrade teacher competencies in interpreting and utilizing test results is widespread (for example, Dunn, 1969; Fleming, 1971; Rost, 1973).

In most cases, parents are also inexperienced in the area of interpreting test results. Wahlstrom, Danley, and Raphael (1977) found most educators reluctant to present "raw results" to parents due to the perception that parents are unable to interpret them properly. It was felt that parents would place too much emphasis on the scores. Others who have expressed concern about overemphasis of test scores include Anastasi (1976), Backman (1976), and Brown (1976).

Presentation of Results to Parents and Students

One topic that has received some attention is the presentation of test score results to students. Since test results contain information of potential benefit it is important that students receive their results in a complete and fully understandable way. Anastasi (1976) indicated a preference for the use of broad levels of performance and qualitative descriptions over numerical scores. Further, her feelings are that test scores "should be accompanied by interpretative explanations by a professionally trained person" (p. 56). Backman (1976) also recommended reporting bands instead of numerical scores to reduce the chance of overemphasizing small differences in test scores.

The incorporation of background information is considered to be important in reporting student test scores as well as in the interpretation of the results of standardized tests (Performance printouts for parents, 1974). For example, a computer reporting system in Memphis allows teachers to include anectodal information about the students in their reports to parents. Teachers select, from a list of statements stored on a computer, those which apply to each student. The printouts

of the statements are sent home after the report cards to provide parents with descriptive information about their child's performance and work habits.

Use of Computer Technology to Report Test Scores

The growth of computer technology has simplified the task of scoring tests and preparing reports. The speed and accuracy of the computer allows the use of various scoring keys with a single test to provide analyses beyond those usually provided (Baker, 1971). Also, computer scoring allows publishers to provide much more information than is currently provided and to provide it in such a way as to help teachers diagnose learning problems (Lewis, 1977).

Nichols and Knopf (1977) listed several advantages of computerized score interpretations beyond those mentioned above. Such systems are faster and less expensive than individual systems. They can be used by persons not trained in test interpretation. Finally, they are less subject to misinterpretation when read by different people than raw or standard scores.

Most available computer scoring systems do not, however, take full advantage of the available options. Many scoring systems provide students with only reports of items missed and identification of the correct responses for those items (Furlong & Miller, 1978). As an example of what is possible, Furlong and Miller (1978) described a reporting system which provides individual reports of (1) an individual's performance relative to other students taking the test, (2) incorrectly answered items and correct responses to those items, (3) the objectives to which incorrectly answered items refer, and (4) if the teacher desires, alternative materials which may be used for further study. The system allows

teachers to receive summaries of performance by item and by objective. The report also summarizes performance by taxonomic level of objectives. Such a system would seem useful for reporting criterion-referenced test score information at the classroom level.

It is clear that computer-generated reports and interpretations offer a promising alternative to traditional reports. A number of the larger test publishing companies are now making considerable use of some of these alternatives. They allow for a variety of scoring schemes, matching of report formats to audience needs, and a reduction in the number of score misinterpretations.

Prerequisites for Desirable Criterion-Referenced Test Score Reporting Systems[1]

A test score report which provides needed information to several groups of interested parties is the product of much research and developmental work. Several activities must occur before high quality reports can be prepared. The activities are listed below:

- Specification of Information Needs,
- Building a Testing Program Consistent with Needs,
- Identification of Audiences and Their Levels of Testing Knowledge,
- Proper Test Selection,
- Proper Test Construction.

The purpose of this section is to briefly discuss these necessary antecedents to the preparation of appropriate reports.

---

[1]It seems clear also that the prerequisites apply equally well to norm-referenced test score reporting systems.

Specification of Information Needs

A school system should clearly specify the groups who are to be served by a testing program and what (specifically) their information needs are. From there, it is possible for a school district to formally state its purposes. Along the way, school subjects, course objectives, and grade levels which will be involved in testing should be specified.

Building a Testing Program Consistent with Needs

Two points are of concern. A test characteristic mentioned by Popham (1978a) is an adequate number of items per measured behavior. Although the number of items desired is not usually specified, a general idea of the relative emphasis to be placed on each domain of content should be available. The other consideration is the scope of the program. A good testing program should provide data in relation to (at least) the most important information needs. It is important, however, that each testing situation include enough items per objective to yield reliable measurement of the objectives of interest at that time without requiring excessive amounts of testing time. It is necessary, therefore, to design a testing program which (1) provides reliable and valid objective score information and (2) presents the data in a manner that will not confuse the audiences or overload them with information which is either too specific or too general for their purposes.

Identification of Audiences and Their Levels of Testing Knowledge

Decisions must be made about the types of information needed and the people who will receive the information. The lack of sophistication of teachers and parents in the field of tests and measurements has been discussed previously. This should be considered when decisions determining the manner in which information will be presented to the various audiences are made. Specific statements of reporting goals at this stage can ease the burden of test selection and dissemination of results.

Proper Test Selection

Test selection cannot be handled appropriately unless information needs are clearly articulated. The task is to identify the available tests which come closest to matching:

—the curricular emphases, and

—the informational requirements of a school program.

Once a limited set of potentially suitable tests has been identified, a detailed evaluation of each test can be carried out. One of the available sets of guidelines for test evaluation will be beneficial (APA, 1974; Hambleton & Eignor, 1978; Popham, 1978a).

Proper Test Construction

Another essential consideration is that of test development. On some occasions individuals will want to develop their own criterion-referenced tests. Such a situation necessitates the availability of staff with both test development training and the time to do the job (Hambleton & Eignor, 1979).

Summary

The purpose of this section has been to emphasize the importance of several prerequisites of appropriate reporting systems. Although the fulfillment of these prerequisites will not guarantee the preparation of a high quality report, the failure to meet them will almost certainly insure low quality reports due to inappropriate or insufficient data. The prerequisites are therefore necessary but not sufficient to insure the preparation of high quality reports of test results. The characteristics of reporting systems which do meet the needs of the several audiences interested in receiving test results and interpretations are considered in the next section.

Characteristics of Desirable Test Score Reporting Systems

In this section elements and options are discussed which should be available in reports of criterion-referenced test scores. Many norm-referenced test score reporting systems will include some of the same information. A logical analysis of the potential of criterion-referenced tests, current uses of the tests, and information needs of various audiences were used to generate the recommendations offered in this section. Four audiences are addressed:

- Teachers,
- Parents and Students,
- Building Administrators, and
- District Administrators.

Table 1 provides a listing of the four audiences, the information to be reported, and the rationale for providing the information. Several of the information needs require explanation beyond that provided in the table. These needs are noted with a numerical superscript. Explanations are found in Appendix A.

Insert Table 1 about here

Table 1

Audiences Desiring Test Results, Their Information Needs and Examples of Uses of the Information Provided

Audience	Information Needs	Use of the Information
Teachers	Master list of objectives tested.	Provide general comparison of test and curricular match
	Information keying items to objectives and objectives to clusters.	Provide specific comparisons of class activities and test.
	Individual student data by objective including raw scores and cut-off scores.	Identification of specific individual deficiencies and the degree of remediation necessary
	Individual student data by objective cluster.	Identification of general areas of individual deficiencies.
	Diagnostic statements of errors of non-masters.	Aid the design of instructional activities to upgrade performance.
	Performance of individuals on previous tests of the same or related objectives.	Identify trends in individual strengths and weaknesses.
	Identification by objective of all students who were classified as masters and those classified as non-masters.	Devise grouping patterns for new instruction and/or remediation.
	Summary class data for each objective.	Identification of specific instructional and/or curricular deficiencies
	Identification of objectives on which performance of the class was low.	Self-evaluation of instruction and determination of needs for group remediation.
	Summary class data for each cluster of objectives.	Identification of general areas of instructional and/or curricular deficiencies.

Table 1 (continued)

Audience	Information Needs	Use of the Information
	Previous performance of the class on the same or related objectives.	Identify trends in class performance.
	Previous performance of students in classes taught by this teacher on the same or related objectives.	Identify trends in effectiveness of curriculum and/or instruction.
	Performance of other classes at the same instructional level in the system.	Determine performance of the class relative to performance in the system.
	Performance of other classes at the same instructional level in the state or nation.	Determine performance of the class relative to state or national performance.
Parents and Students[1]	Performance on clusters of objectives.	Provide general overview of performance.
	Identification of specific objectives on which performance is low.	Determine specific deficiencies.
	Inclusion of sample items from non-mastered objectives.	Clarification of skills to be mastered.
	Identification of trends of performance across tests or subtests.	Identification of strengths and weaknesses in broad areas of performance.
	Performance from previous tests on the same or related objectives.	Identification of trends of improvement or decline.
	Performance relative to other students in the same class.	Determine relative standing in the class.
	Performance relative to other students at the same instructional level in the system.	Determine relative standing in the system.
	Performance relative to other students at the same instructional level in the state or nation.	Determine relative performance as compared to a national sample.
	Performance in relation to aptitudes.	Determine if student is performing to his/her potential.

Table 1 (continued)

Audience	Information Needs	Use of the Information
	Narrative diagnostic and interpretive reports to supplement numerical summaries.	Reduce misunderstanding of scores, provide alternative views of the data, identify areas needing attention.
	Statement from an official of the school system (if desired by system).[2]	Explain some aspect of the testing program.
	Comments from student's teacher.[3]	Provide background information to enhance interpretation of scores.
Building Administrators	Summaries of subtest performance for each classroom	Identification of classes which may need specific remediation.
	Summaries of subject performance for each classroom.	Identification of classes which may need general remediation. Determine the need for added personnel or in-service in a subject.
	Summaries of subtest performance by grade for the school.	Identify trends of performance.
	Identification, for each subtest, of clusters of objectives on which performance was low.	Indicate the need for curricular and/or instructional revision.
	Summaries of subject performance by grade for the school.	Identify subjects in need of curricular and/or instructional revision or increased resources.
	Summaries of subject performance by grade on previous tests of the same objectives.	Identify areas of improvement of decline.
	Summaries of student performance on key objectives.[4]	Monitor progress on school or district priorities.
	Summaries of student performance by grade for other district schools.	Comparison and identification of specific strengths and weaknesses. Identification of trends in the district.

Table 1 (continued)

Audience	Information Needs	Use of the Information
	Summaries of subject performance by grade for other district schools.	Comparison with other schools on a general basis. Identification of general performance trends.
	Master list of objectives and percentage of students classified as masters in each school and the district.	Reference and comparison.
	Individual permanent record labels.	Student files.
	Performance by grade relative to state or national norms for each subtest (optional).	Determine relative standing of classes of students.
District Administrators	Summaries of subtest performance in each grade by school.	Determine achievement levels.
	Summaries of subject performance in each grade by school.	Determine schools in need of additional resources (financial or special personnel).
	Summaries of subtest performance in each grade for the district.	Public release.
	Summaries of subject performance in each grade for the district.	Public release.
	Identification, for each subtest, those schools in which performance was low.	Determine in-service needs.
	Summaries of subject performance by grade on previous tests of the same objectives.	Identify trends of improvement or decline in the district.
	Master list of objectives, number of items per objective, cut-off scores, and percentage of students in the district exceeding the cut-off score.	Reference.
	Summaries of student performance on key objectives.[4]	Monitor progress on school or district priorities.

Table 1 (continued)

Audience	Information Needs	Use of the Information
	"Split" summaries of subject performance by designated subgroups (race, sex, etc.).	Public release, reports to government officials.
	Normative data of subtest performance relative to the state or nation.	Comparison of achievement with other districts.
	Computer tapes containing "raw" data of student performance.	Research studies within the system.

Four matters that arise when designing or evaluating reporting systems are:

- Physical features
- Reporting normative information
- Choice of cut-off scores
- Test score interpretations.

Each will be considered below.

Physical Features. The page size of reports can be a problem. Small page sizes are easy to lose or damage. Large page sizes, which are common, are cumbersome and hard to store in standard folders or notebooks. Therefore, the option to print reports on standard 8½" x 11" paper would be desirable. Two potential disadvantages of the 8½" x 11" page size are, however, the reduced amount of information which can be printed on a page and the possibility of decreased readability of the report forms due to the use of small print in order to include more information on the page.

Regardless of page size, each page should list the audience to receive the report, the test administration date, the information included on that page, and the examinee or group of examinees covered by the information. An address to write to or a phone number to call for clarification of confusing points is also desirable on school and district administrator report forms. Reports should arrive in the format in which they will be distributed. That is, school personnel should not be required to fold, cut or paste reports for the audiences served by the testing program.

Whenever possible, all information pertaining to one test or subtest should be included on a single page. This eliminates the need for referring back and forth between pages to make comparisons. Attempts to provide all data on one page should not, however, forsake legibility; sufficient space should be allowed between columns and rows of scores to allow easy reading. Reports which have alternating rows or columns of shaded and nonshaded background facilitate legibility. Narrative passages should be within a page of the tables to which they refer if they cannot be included on the same page. Not only does this keep related information together, but it also separates tables of numbers from one another which improves the ease of reading the report.

Reporting Normative Information. It is highly desirable to report norm-referenced and criterion-referenced test score information, especially when it is derived from the same test, on the same page or pages of print-out sheets. Figure 1 provides a good illustration of this point. To do otherwise invites confusion which will lead to misinterpretation. Whenever possible, error bands should be reported along with the test score information.

Insert Figure 1 about here

Choice of Cut-off Scores. Permitting choice of cut-off scores on the objectives by teachers, schools, or districts is a feature that will enhance the utility of a testing program and the associated reporting system. Since schools will atach different levels of importance to objectives, it is reasonable

District: Pine Valley
School: George Washington
Class: Mr. Barber (3A)

PROGRAM EVALUATION REPORT
INDIVIDUALIZED CRITERION-REFERENCED TESTS
CLASS SUMMARY
-READING-

Batch Number: 431
Pretest Date: 9/15/79
Posttest Date: 5/12/80
Page: 4

NORM-REFERENCED INFORMATION					STUDENT INFORMATION		CRITERION-REFERENCED INFORMATION													
							OBJECTIVES/STANDARDS (%)											COMMON OBJECTIVES		
Admin.	Scaled Score	GE	%ile	NCE	Name	ID	1 80	2 80	3 100	4 80	5 70	6 70	7 70	8 80	9 80	10 80	Number Mastered	Number	Percent Mastered	Growth: % of Potential
Pre	215	2.1	31	40	Abernathy, John	146	30	90+	60	20	40	10	50				1		14	
Post	270	3.4	46	48			50	100+	80	60	30	40	70+	40	40	80+	3	7	29	17
Pre	250	2.8	43	46	Adams, Cathy	147	80+	40	80	80+	70+	60	60	40	30		3		33	
Post	310	4.2	54	52			80+	60	100+	80+	90+	100+	80+	60	60	80+	7	9	67	50
Pre	185	1.5	19	32	Baker, Maria	148	60	50	70	40	50	60	30	20			0		0	
Post	235	2.7	41	45			80+	60	90	60	50	70+	40	30	60		2	8	25	25
Pre	230	2.4	39	44	Booth, Derek	149	50	100+	80	60	50	100+	60	60	70	20	2		20	
Post	300	4.0	51	50			80+	70	100+	80+	60	100+	60	60	90+	60	5	10	50	38
Post	250	2.8	43	46			80+	80+	100+	40	20	0	100+	100+	90+	50	6	8	75	60

SUMMARY STATISTICS
(Number of Students=31)

Variable	Admin.	Average	Standard Deviation	Mean Change
Scale Score	Pre	224.1	50.2	
	Post	285.2	57.3	61.1
Grade Equivalent Score	Pre	2.3	.7	
	Post	3.7	.8	1.4
NCE Score	Pre	36.0	20.3	
	Post	47.2	22.6	11.2
Number of Objectives Mastered	Pre	1.7	.9	
	Post	5.9	1.5	4.2
Common Objectives				
Number	--	9.2	1.7	--
% Mastered	Pre	19.3	10.1	
	Post	49.0	16.2	29.7
Growth: % of Potential	--	37.3	17.9	--

Percent of Students in Each Quarter of the National Norm Distribution

Admin.	Quarter: Low (0-25)	Lower Middle (26-50)	Upper Middle (51-75)	High (76-100)
Pre	36	35	17	12
Post	21	30	36	13

Figure 1. A sample report with both norm-referenced and criterion-referenced test score information. The authors are grateful to Educational Progress Corporation for permission to reproduce a modified version of one of their reporting forms.

to assume that students would be expected to perform better on some objectives than others. School systems could receive instructions on procedures which could be used to choose an appropriate cut-off score for each objective. Alternately, it is possible to provide school systems with a list of objectives and (say) three possible cut-off scores for each objective. Cut-off scores can then be chosen to reflect the level of importance placed on the objectives measured at a certain grade level in the system.

Test Score Interpretations. Many of the tests which are currently called criterion-referenced tests are more accurately described by the term objective-referenced tests (Hambleton & Eignor, 1978). The difference between the two types of tests is an important one because it impacts on the kinds of test score interpretations which are permissible. An objective-referenced test is one in which test items are keyed to behavioral objectives. Scores on such a test reflect examinee performance on those particular items which make up the test. A criterion-referenced test, on the other hand, is composed of groups of items which represent samples from well-defined content or behavior domains. Such a test allows an examinee's score to be interpreted not only in relation to the items on the test, but also in relation to the entire content or behavior domains sampled by the test. It is the latter interpretation that is most often desired (so much so, that often such interpretations are made whether they are justified or not!). A test score report should include a section describing the types of inferences from test scores that are possible. Failure to provide such information invites over- or under-interpretation of the scores.

Guidelines for Designing and Evaluating Test Score Reporting Systems

A rating system which can be used in designing or evaluating criterion-referenced test score reporting systems was designed based on information provided in Table 1. The system can be used to specify the types of data required to meet practitioners' previously specified informational needs and then to either design a system to meet those needs or to evaluate the reporting systems which accompany the tests which are being considered. Practitioners first determine the types of data required and the importance of each type of data. Reporting systems are then evaluated in terms of the extent to which they meet the requirements of the local school system.

A set of 69 questions organized around seven major categories of concern in relation to reporting systems is presented in Figure 2. The seven categories are:

A. Information about objectives

B. Information at the item level

C. Information at the objective level

D. Information at the cluster of objectives or subtest level

E. Subject summaries

F. Specialized services

G. General services

Insert Figure 2 about here

We have also suggested questions in Figure 2 which are often of concern to teachers, to parents and students, to building administrators,

Figure 2. A set of questions about test score reports. (T=teacher; P=parent and student; B=building; D=district.)

In order to meet our informational needs, should (does) the reporting system. . .	T	P	B	D
A. ...provide information about the objectives?	X[1]	X	X	X
1. ...list all objectives tested?	X	X	X	X
2. ...list the number of items per objective?	X	X	X	X
3. ...identify the items measuring each objective?	X			
4. ...list the cut-off score for each objective?	X	X	X	X
B. ...provide information at the item level?	X			X
1. ...identify items answered correctly?	X			
2. ...identify responses chosen for incorrectly answered items?	X			
3. ...provide a data tape or punched cards of all examinees and item responses?				X
C. ...provide information at the objective level?	X	X	X	X
1. ...report individual scores by objective?	X	X		
2. ...report names of masters and non-masters by objective?	X			
3. ...provide diagnostic statements about performance on non-mastered objectives?	X	X		
4. ...report classroom average for each objective?	X	X	X	
5. ...report number or percent masters for each objective?	X	X	X	X
6. ...report objectives on which class performance was low?	X		X	
7. ...report average state or national performance by objective?	X	X	X	X
8. ...identify mastered and non-mastered objectives?		X		
9. ...provide examples of items measuring non-mastered objectives?		X		
10. ...report past performance of students on the same or related objectives?	X	X		
11. ...report past performance of the class on the same or related objectives?	X		X	
12. ...report the number of students tested in the classroom for each objective?	X		X	
13. ...report average school performance by objective?	X	X	X	X
14. ...report the number of students tested in the school for each objective?			X	X
15. ...report average district performance by objective?			X	X
16. ...report the number of students tested in the district for each objective?				X
17. ...report past performance of the school on the same or related objectives?			X	X
18. ...report past performance in the district on the same or related objectives?				X
19. ...report objectives on which school performance was low?			X	X
20. ...report objective scores as raw scores?	X	X	X	X
21. ...report objective scores as percent scores?	X	X	X	X
22. ...report frequency distribution of percent of objectives mastered?	X		X	X

[1]"X" designates possible informational needs.

Figure 2. (continued)

In order to meet our informational needs, should (does) the reporting system. . .	T	P	B	D
D. ...provide data by cluster of objectives or subtest?	X	X	X	X
1. ...report individual scores on each objective cluster?	X	X		
2. ...report average class score on each objective cluster?	X	X	X	
3. ...report average school score on each objective cluster?	X	X	X	X
4. ...report average score of students in the state or nation on each objective cluster?	X	X	X	X
5. ...report number or percent of objectives mastered by individuals for each objective cluster?	X	X		
6. ...report average number or percent of objectives mastered in the class for each objective cluster?	X	X	X	
7. ...report average number or percent of objectives mastered in the school for each objective cluster?	X	X	X	X
8. ...report the average score in the district for each objective cluster?			X	X
9. ...report the average number or percent of objectives mastered in the district for each objective cluster?			X	X
10. ...report past classroom performance on each objective cluster?	X		X	
11. ...report past school performance on each objective cluster?	X		X	X
12. ...report past district performance on each objective cluster?			X	X
13. ...identify objective clusters for which overall class performance was low?	X		X	
14. ...identify objective clusters for which overall school performance was low?			X	X
15. ...identify objective clusters for which overall district performance was low?				X
16. ...report normative information for objective clusters?	X	X	X	X
E. ...report subject summaries?	X	X	X	X
1. ...report individual student scores for each subject?	X	X		
2. ...report average classroom scores for each subject?	X	X	X	
3. ...report average school scores for each subject?	X	X	X	X
4. ...report average district scores for each subject?		X	X	X
5. ...report past classroom performance for each subject?	X	X	X	
6. ...report past performance of individual students for each subject?	X	X		
7. ...report past performance of the school for each subject?	X	X	X	X
8. ...report past performance of the district for each subject?			X	X
9. ...report normative information for each subject?	X	X	X	X

Figure 2. (continued)

In order to meet our informational needs, should (does) the reporting system. . .	T	P	B	D
F. ...provide specialized services?	X	X	X	X
1. ...include a standard statement from the superintendent or other district official?		X		
2. ...include teacher comments about the student?		X		
3. ...highlight data on selected "key" objectives?	X	X	X	X
4. ...provide "split" data on specific subgroups (race, special programs, etc.)?			X	X
5. ...provide individual reports for student record files?			X	
6. ...report performance in relation to aptitudes?	X	X		
G. ...provide general services?	X	X	X	X
1. ...print all reports on 8½" x 11" paper?	X	X	X	X
2. ...identify, on each page, the audience for whom the report is intended, the type of information included, and the test date?	X	X	X	X
3. ...report normative information as a band or as a number with an error term?	X	X	X	X
4. ...include all information relating to one test or subtest on the same page?	X	X	X	X
5. ...devote a section of the report to the generalizability of the scores?	X	X	X	X
6. ...provide suitable forms and directions for hand-scoring?	X			
7. ...provide reports in a format which does not require further preparation (folding, cutting)?	X	X	X	X
8. ...provide a telephone number to call for aid in interpreting the reports?			X	X
9. ...provide narrative explanations of data?	X	X	X	X

and to district administrators. The useful questions will, however, vary from one situation to the next and so the designations in Figure 2 are not absolutes.

How might someone evaluate the test score reporting component of one or more criterion-referenced tests using the questions in Figure 2? We have prepared an evaluation form and a scoring system[1] which should be useful. An excerpt from the form is presented in Figure 3.

Insert Figure 3 about here

The figure contains a portion of a hypothetical evaluation of the teacher reports provided by three criterion-referenced tests. It was completed according to the directions below:

1. Read the phrase "In order to meet our informational needs, should the reporting system..." and the major section heading.
2. Rate the importance of each major section of questions on a scale from 0 to 4. Zero indicates no importance; 4 indicates that the information is very important. Place the "importance rating" in the first column to the right of the question. In the example, providing information about the objectives was viewed by the rater as somewhat important (2), item data was unimportant (0), and receiving objective data was very important (4).
3. For each major section receiving an importance rating greater

[1]Space does not permit the long evaluation form to be reproduced here. It can be obtained by writing to the authors.

I. Teacher Reports	Importance	System #1	System #2	System #3
A. *Provide information about the objectives?*	2	.67	2.0	1.33
1. List all objectives tested?	✓	✓	✓	✓
2. List the number of items per objective?	✓		✓	✓
3. Identify the items measuring each objective?				
4. List the cut-off scores for each objective?	✓		✓	
B. *Provide information at the item level?*	0			
1. Identify items answered correctly?				
2. Identify response chosen for incorrectly answered items?				
C. *Provide information at the objective level?*	4	2.0	2.67	2.67
1. Report individual scores by objective?	✓	✓	✓	✓
2. Report names of masters and non-masters by objective?	✓			
3. Provide diagnostic statements about performance on non-mastered objectives?				
4. Report classroom average for each objective?	✓	✓	✓	✓
5. Report number of percent masters for each objective?	✓		✓	
6. Report objectives on which class performance was low?				
7. Report average state or national performance by objective?				
10. Report past performance of students on the same or related objectives?				
11. Report past performance of the class on the same or related objectives?				
13. Report the number of students tested in the classroom for each objective?	✓			✓
14. Report average school performance by objective?	✓	✓	✓	✓
OVERALL TOTAL		11.4	13.8	17.0

Figure 3. A portion of an evaluation of the teacher reports provided by the reporting systems which accompany three criterion-referenced tests.

than zero, read the numbered questions which follow it. Precede each question with the same phrase as in step one. Place a check (✓) in the box immediately to the right of all questions answered affirmatively. (In the example, the following questions were answered in the affirmative: A.1, A.2, A.4, C.1, C.2, C.4, C.5, C.13, C.14.)

4. Place the names of the tests under review in the proper places.
5. Answer only the numbered questions which received a "✓" in step three. Place a "✓" in the appropriate boxes for questions which are answered affirmatively. (For example, with the first test, questions A.1, C.1, C.4, and C.14 were answered in the affirmative.)
6. Within each section, count the number of "✓" marks given to the test.
7. Of the questions of concern, calculate the proportion of questions which are answered in the affirmative.
8. Multiply the proportion by the "importance rating" for each section.
9. Place the resulting numbers in the boxes which are on the same lines as the major section names and under the appropriate reporting system.
10. Repeat steps 5 through 9 for the reporting system which accompanies each of the tests under consideration.
11. After all reports have been evaluated, sum the numbers recorded for each reporting system. The one with the highest total is the one which comes closest to satisfying important informational needs.

There are many possible methods of weighting the major categories and questions presented in the evaluation system. Practitioners are urged to use the one which appeals most strongly to them. The advantage to the system presented here is that it weights all sections equally regardless of the number of questions checked. The maximum possible score a reporting system can obtain for a section is the importance rating placed on that section. Another, and even simpler descriptive statistic for evaluating the effectiveness of a criterion-referenced test score reporting system would be the percent of times desirable features are provided by the reporting system. For example, if 25 features are desired and 20 of them are provided by a reporting system, the descriptive statistic "80%" could be used to compare the system to others.

Conclusions

Although research during the last decade fostered many advances in criterion-referenced testing technology, one area which was not extensively researched was the reporting of the results of the tests. Since the test results are used to make a variety of decisions, some of which have long-term effects on students, it is important that research be conducted into the best approaches for reporting test score information. That the National Institute of Education is beginning to fund research in the area is a clear sign of the recognition of the need for technological advances.

In this paper, we have reviewed the literature related to reporting test results, discussed the characteristics of high quality reports, and presented an evaluation system which can be used to determine data necessary to meet informational needs and to design a reporting system to provide that data, or which can be used in conjunction with test evaluation guidelines as an overall evaluation of testing systems under consideration. On the matter of guidelines, it was not our intention to offer a common list of requirements for all reporting systems. Instead, a list was presented which identified a large number of characteristics of reporting systems which are appropriate in many situations. Practitioners may choose, from the list, those which apply to their situation. Reporting systems should then be evaluated by the practitioner in light of local information and financial priorities.

The evaluation system presented here is an initial attempt to develop a flexible system for development or evaluation of reporting systems. While the evaluation system is the result of careful analysis

of informational needs of users and the potential of criterion-referenced tests, it has not been extensively field-tested. We expect that some questions may be deleted and others added as the system is field-tested. It is, however, our expectation that the evaluation system will provide a useful tool for practitioners developing and/or evaluating reporting systems and for researchers interested in the areas of test score reporting and interpretation.

Since the reporting of test score information represents an interface between measurement experts and several audiences with limited knowledge about the field of tests and measurements, the possibility for miscommunication is high. Further research is needed to reduce this possibility. Among the questions to be answered are the following:

1. How should practitioners determine and prioritize their informational needs?
2. How well do present commercial criterion-referenced test score reporting systems meet the needs of consumers? What do the various audiences discussed in this paper perceive as the strengths and weaknesses of the reports? Are they satisfied with the quantity and quality of current reports? What report formats are most effective at meeting their needs?
3. What types of instructional aids are needed to increase the likelihood of correct test score interpretations? What basic information do the various audiences need about the results of criterion-referenced tests and what are the best ways of providing the information?

4. How can information from norm-referenced or criterion-referenced tests be combined in a reporting system to maximize the usefulness of the test scores?
5. Do new types of measurement scales (such as those provided through the use of latent trait theory) have the potential for conveying more information about examinee performance?
6. How should group information be provided? This paper has considered mainly reports based on individual student data. The use of tests to meet accountability demands and for program evaluations necessitates research in the area of reporting group-based information.

Although criterion-referenced tests are in wide use and much of the technology is in place to develop tests and to validate descriptions and decisions arising from the uses of test scores, the tests have not reached their full potential in those districts which use them. Continuing improvement in the quality of test score reports and advances in other related technological areas are necessary to ensure that criterion-referenced tests meet the needs of educators.

References

Ahmann, J. S. Basic issues concerning competency-based testing. In R. B. Ingle, M. R. Carroll, & W. J. Gephart (Eds.), Assessing student competence in the public schools. Bloomington, IN: Phi Delta Kappa, 1978, pp. 78-89.

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Anastasi, A. Psychological testing (4th ed.). New York: MacMillan, 1976.

Backman, M. E. Reporting test results for guidance and instructional purposes. In Guide for School Testing Programs, Washington: National Council on Measurement in Education, 1976.

Baker, F. B. Automation of test scoring, reporting, and analysis. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971, pp. 202-234.

Bradley, R. W. Person-referenced test interpretation: A learning process. Measurement and Evaluation in Guidance, 1978, 10, 201-210.

Brown, F. G. Principles of educational and psychological testing (2nd ed.). New York: Holt, Rinehart, and Winston, 1976.

Dunn, S. S. Helping teachers to make better use of test results. In K. Ingekamp (Ed.), Developments in educational testing. London: University of London Press, 1969.

Fisher, T. H. The assessment of student competencies in Florida's public schools. In R. B. Ingle, M. R., Carroll, & W. J. Gephart (Eds.), The assessment of student competence in the public schools. Bloomington, IN: Phi Delta Kappa, 1978, pp. 24-38.

Fleming, M. Standardized tests revisited. The School Counselor, 1971, 19, 71-72.

Furlong, F., & Miller, W. DIAGNOSE: Computer-based reporting of criterion-referenced test results. Educational Technology, 1978, 18, 37-39.

Gardner, E. F. Interpreting achievement profiles: Uses and warnings. Journal of Research and Development in Education, 1977, 10, 51-63.

Goslin, D. A. Teachers and testing. New York: Russel Sage Foundation, 1967.

Hagen, E. Use and abuse of tests in education. In R. M. Bossone and M. Weiner (Eds.), Proceedings of the National Conference on Testing: Major Issues. University of New York, November 1977. (ERIC Document Reproduction Service No. ED 152 814.)

Hambleton, R. K., & Eignor, D. R. Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70 (2nd ed.). Amherst, MA: School of Education, University of Massachusetts, 1979.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Kirby, J. H., Culp, W. H., & Kirby, J. MUST: Manual for the users of standardized tests. Bensonville, IL: Scholastic Testing Service, 1973.

Lawson, J. H., & Ward, A. W. Use of tests in program evaluation. In Guide for School Testing Programs. Washington: National Council on Measurement in Education, 1976.

Lewis, B. Testing: A parent's point of view. In R. M. Bossone and M. Weiner (Eds.), Proceedings of the National Conference on Testing: Major Issues. University of New York, November 1977. (Eric Document Reproduction Service No. ED 152 814.)

Millman, J. Reporting student progress: A case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 226-230.

Millman, J. Strategies for constructing criterion-referenced assessment instruments. Paper presented at the eighth annual conference on large-scale assessment, Boulder, Colorado, 1978.

Nichols, M. P., & Knopf, I. J. Refining computerized test interpretations: An in-depth approach. Journal of Personality Assessment, 1977, 41, 157-159.

Page, E. B. Testing: What in the world are we arguing about. In R. M. Bossone and M. Weiner (Eds.), Proceedings of the National Conference on Testing: Major Issues. University of New York. November 1977. (ERIC Document Reproduction Service No. ED 152 814.)

Performance printouts for parents. Nation's Schools and Colleges, September 1974, 1, 31-32.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978. (a)

Popham, W. J. Practical criterion-referenced measures for intrastate evaluation. Educational Technology, 1978, 18, 19-23. (b)

Rost, P. Useful interpretation of standardized tests. Clearing House, January 1973, 47, 319-320.

Stetz, F. P. Providing relevant data for decision-making purposes. The Elementary School Journal, June 1978, 78, 220-225.

Wahlstrom, M. W., Danley, R. R., & Raphael, D. Measuring achievement at the primary and junior levels: An analytical review of test instruments used in evaluating pupil achievement and of communicating results to parents. Toronto, Ontario: Ontario Institute for Studies in Education, 1977.

Wahlstrom, M. W., Regan, E., & Jones, L. L. An analysis of teacher beliefs in relation to procedures for assessing student achievement. Toronto, Ontario: Ontario Institute for Studies in Education, 1978.

Appendix A

Notes to Accompany Table 1

1. It should be noted that the data provided to parents is less specific than the data provided to teachers. While teachers need specific information on every objective in order to devise instructional prescriptions, parents and students may only require information about performance on clusters of related objectives. A report of an arithmetic computation test might include information about the student's overall performance in addition, multiplication, subtraction, and division. Further subdivisions such as sutraction of whole numbers, subtraction of decimals, subtraction of fractions, etc., might provide too many sets of data for the parents. It would be better to provide an overall performance appraisal for subtraction and then identify areas which need further work.

   For example, a child might answer 28 out of 35 subtraction problems correctly, but only correctly answer 3 out of 7 questions dealing with the subtraction of fractions. The report to the parent would say that the student had answered 80% of the items which related to subtraction correctly, but that subtraction of fractions was an area where performance was low. Statements of objectives and example items for those areas which show less than adequate achievement should be included.

2. Communication between schools and parents is often neglected. When reports of test results reach the parents they are often unaware of the purpose or scope of the testing program. A short statement from the superintendent or some other official would enhance the acceptance and understanding of the program and the scores reported.

3. Teachers are in possession of a wealth of information which could enhance student and parent understanding of test score reports. Teachers could receive a coded list of statements concerning classroom activities. Teachers could select, from the list, those statements which apply to each student. The codes could be recorded on the student's answer sheet. A computer program could then include the statements in the individual reports. Statements could range from identification of objectives which have not yet been taught to statements pertaining to an individual's interest in a given subject area.

4. Often a school or school district will choose a small number of key objectives on which to concentrate in a given year. The option should exist for a number of objectives (2-3 per subject area) to be classified as key objectives. Data on these key objectives should be presented to building administrators and to system administrators.